Observing many researchers using the same data and hypothesis reveals a hidden universe of data analysis

Breznau, Nate; Rinke, Eike Mark; Wuttke, Alexander; Jaeger, Bastian

DOI:
10.31222/osf.io/cd5j9

Publication date:
2021

Document Version
Early version, also known as pre-print

Observing Many Researchers using the Same Data and Hypothesis Reveals a Hidden Universe of Data Analysis

~

Principal Investigators:

Nate Breznau, University of Bremen, breznau.nate@gmail.com
Eike Mark Rinke, University of Leeds, E.M.Rinke@leeds.ac.uk
Alexander Wuttke, University of Mannheim, alexander.wuttke@uni-mannheim.de

Crowdsourced Researcher Co-Authors:

Muna Adem[21], Jule Adriaans[10], Amalia Alvarez-Benjumea[35], Henrik K. Andersen[6], Daniel Auer[37], Flavio Azevedo[62], Oke Bahnsen[37], Dave Balzer[24], Gerrit Bauer[32], Paul C. Bauer[37], Markus Baumann[17], Sharon Baute[56], Verena Benoit[57,32], Julian Bernauer[37], Carl Berning[24], Anna Berthold[57], Felix S. Bethke[42], Thomas Biegert[33], Katharina Blinzler[11], Johannes N. Blumenberg[11], Licia Bobzien[18], Andrea Bohman[52], Thijs Bol[56], Amie Bostic[50], Zuzanna Brzozowska[88,85], Katharina Burgdorf[37], Kaspar Burger[2,55], Kathrin Busch[20], Juan Carlos-Castillo[61], Nathan Chan[49], Pablo Christmann[11], Roxanne Connelly[1], Christian S. Czymara[13], Elena Damian[27], Alejandro Ecker[37], Achim Edelmann[47], Maureen A. Eger[52], Simon Ellerbrock[37], Anna Forke[48], Andrea Forster[56], Chris Gaasendam[27], Konstantin Gavras[37], Vernon Gayle[1], Theresa Gessler[2], Timo Gnambs[29], Amélie Godefroidt[40], Max Grömping[14], Martin Groß[82], Stefan Gruber[59], Tobias Gummer[11], Andreas Hadjar[74], Jan Paul Heisig[73], Sebastian Hellmeier[69], Stefanie Heyne[37], Magdalena Hirsch[73], Mikael Hjerm[52], Oshrat Hochman[11], Andreas Hövermann[15], Sophia Hunger[73], Christian Hunkler[19], Nora Huth[84], Zsófia S. Ignácz[13], Laura Jacobs[53], Jannes Jacobsen[9,5,10], Bastian Jaeger[51], Sebastian Jungkunz[64,57,65], Nils Jungmann[11], Mathias Kauff[36], Manuel Kleinert[25], Julia Klinger[62], Jan-Philipp Kolb[7], Marta Kołczyńska[43], John Kuk[79], Katharina Kunißen[75], Dafina Kurti Sinatra[20], Alexander Langenkamp[13], Philipp M. Lersch[19,10], Lea-Maria Löbel[10], Philipp Lutscher[80], Matthias Mader[72], Joan E. Madia[41], Natalia Malancu[67], Luis Maldonado[44], Helge-Johannes Marahrens[21], Nicole Martin[76], Paul Martinez[90], Jochen Mayerl[6], Oscar J. Mayorga[60], Patricia McManus[21], Kyle McWagner[49], Cecil Meeusen[27], Daniel Meierrieks[73], Jonathan Mellon[76], Friedolin Merhout[63], Samuel Merk[66], Daniel Meyer[62], Leticia Micheli[30], Jonathan Mijs[16,8], Cristóbal Moya[4], Marcel Neunhoeffer[37], Daniel Nüst[78], Olav Nygård[31], Fabian Ochsenfeld[34], Gunnar Otte[24], Anna Pechenkina[86], Christopher Prosser[46], Louis Raes[51], Kevin Ralston[1], Miguel Ramos[58], Arne Roets[12], Jonathan Rogers[39], Guido Ropers[37], Robin Samuel[74], Gregor Sand[59], Ariela Schachter[89], Merlin Schaeffer[63], David Schieferdecker[9], Elmar Schlueter[68], Regine Schmidt[57], Katja M. Schmidt[10], Alexander Schmidt-Catran[13], Claudia Schmiedeberg[32], Jürgen Schneider[82], Martijn Schoonvelde[54], Julia Schulte-Cloos[32], Sandy Schumann[55], Reinhard Schunck[84], Jürgen Schupp[10], Julian Seuring[57], Henning Silber[11], Willem Sleegers[51], Nico Sonntag[75], Alexander Staudt[20], Nadia Steiber[83], Nils Steiner[24], Sebastian Sternberg[26], Dieter Stiers[27], Dragana Stojmenovska[56], Nora Storz[87], Erich Striessnig[83], Anne-Kathrin Stroppe[11], Janna Teltemann[71], Andrey Tibajev[31], Brian Tung[89], Giacomo Vagni[55], Jasper Van Assche[12,27], Meta van der Linden[8], Jolanda van der Noll[70], Arno Van Hootegem[27], Stefan Vogtenhuber[22], Bogdan Voicu[45,77], Fieke Wagemans[38], Nadja Wehl[3,57,72], Hannah Werner[27], Brenton M. Wiernik[81], Fabian Winter[35], Christof Wolf[11], Yuki Yamada[28], Nan Zhang[35], Conrad Ziller[64], Stefan Zins[23], Tomasz Żółtak[43]

Research Assistant:

Hung H.V. Nguyen, University of Bremen hunghvnguyen@gmail.com

[1]University of Edinburgh, [2]University of Zurich, [3]Bamberg Graduate School of Social Sciences, [4]Bielefeld University, [5]German Socio-Economic Panel Survey (SOEP), [6]Chemnitz University of Technology, [7]Destatis, [8]Erasmus University Rotterdam, [9]Free University Berlin, [10]German Institute for Economic Research (DIW Berlin), [11]GESIS - Leibniz Institue for the Social Sciences, [12]Ghent University, [13]Goethe University Frankfurt, [14]Griffith University, [15]Hans-Böckler-Foundation, [16]Harvard University, [17]Heidelberg University, [18]Hertie School, [19]Humboldt University Berlin, [20]Independent research, [21]Indiana University, [22]Institute for Advanced Studies, Vienna, [23]Institute for Employment Research (IAB), [24]Johannes Gutenberg University Mainz, [25]Justus Liebig University Giessen, [26]KPMG, [27]Catholic University Leuven, [28]Kyushu University, [29]Leibniz Institute for Educational Trajectories, [30]Leibniz University Hannover, [31]Linköping University, [32]Ludwig Maximilian University Munich, [33]London School of Economics, [34]Max Planck Society, [35]Max-Planck-Institute for Research on Collective Goods, [36]Medical School Hamburg, [37]University of Mannheim, [38]Netherlands Institute for Social Research, [39]New York University Abu Dhabi, [40]Norwegian University of Science and Technology,

Abstract:

Findings from 162 researchers in 73 teams testing the same hypothesis with the same data reveal a universe of unique analytical possibilities leading to a broad range of results and conclusions. Surprisingly, the outcome variance mostly cannot be explained by variations in researchers' modeling decisions or prior beliefs. Each of the 1,261 test models submitted by the teams was ultimately a unique combination of data-analytical steps. Because the noise generated in this crowdsourced research mostly cannot be explained using myriad meta-analytic methods, we conclude that idiosyncratic researcher variability is a threat to the reliability of scientific findings. This highlights the complexity and ambiguity inherent in the scientific data analysis process that needs to be taken into account in future efforts to assess and improve the credibility of scientific work.

Summary:

Idiosyncratic variability in findings using hypothesis and data offers new insights into scientific uncertainty.

Supplementary:

Replication / Data Repository

Executive Report (some background on the larger project)

Crowdsourced Researcher Survey via Dataverse

Interactive Data Visualization ('Shiny app')

Supplementary Materials (Appendix, Figures, Tables)

[41]Nuffield College, University of Oxford, [42]Peace Research Institute Frankfurt, [43]Polish Academy of Sciences, [44]Pontifical Catholic University of Chile, [45]Romanian Academy, [46]Royal Holloway, University of London, [47]Sciences Po, [48]Stadtberatung Dr. Sven Fries, [49]The University of California, Irvine, [50]The University of Texas Rio Grande Valley, [51]Tilburg University, [52]Umeå University, [53]Université Libre de Bruxelles, [54]University College Dublin, [55]University College London, [56]University of Amsterdam, [57]University of Bamberg, [58]University of Birmingham, [59]Max Planck Institute for Social Law and Social Policy, [60]University of California, Los Angeles, [61]University of Chile, [62]University of Cologne, [63]University of Copenhagen, [64]University of Duisburg-Essen, [65]Zeppelin University, [66]University of Education Karlsruhe, [67]University of Geneva, [68]University of Giessen, [69]University of Gothenburg, [70]University of Hagen, [71]University of Hildesheim, [72]University of Konstanz, [73]WZB Berlin Social Science Center, [74]University of Luxembourg, [75]University of Mainz, [76]University of Manchester, [77]Lucian Blaga University of Sibiu, [78]University of Münster, [79]University of Oklahoma, [80]University of Oslo, [81]University of South Florida, [82]University of Tübingen, [83]University of Vienna, [84]University of Wuppertal, [85]Austrian Academy of Sciences, [86]Utah State University, [87]Utrecht University, [88]Vienna Institute of Demography, [89]Washington University in St. Louis, [90]Western Governors University.

Organized scientific knowledge production has institutionalized checks such as editorial vetting, peer-review, and methodological standards(*1*, *2*) to ensure that findings are independent from the characteristics or predispositions of any single researcher. These procedures should generate inter-researcher reliability offering consumers of scientific findings assurance that they are not arbitrary flukes but that other researchers would generate similar findings given the same data. Recent meta-science challenges this assumption. Attempts to reproduce other researchers' computations, make alternative model specifications and conduct replications with new data often lead to results differing substantially from original studies(*3*, *4*). In response, scientists are paying more attention to institutionalized and procedural sources of bias, for example the interplay of structural incentives and psychological predispositions(*5*–*7*).

We argue that noise in the research process is also a product of the researchers themselves when faced with a universe of analytical flexibility. We highlight this finding based on a comparatively well-defined data analysis task for highly skilled and accuracy-motivated social scientists testing the same hypothesis with the same data. The resulting idiosyncratic unexplained variability suggests there is great complexity and uncertainty inherent to the process of data analysis and conclusion drawing that exists independently of any perverse or biasing structural incentives.

Researcher Variability

What appear as seemingly trivial steps in data analysis(*8*, *9*) may introduce variability and influence research outcomes to a limited extent(*10*). Yet as these steps accumulate across an entire workflow, outcomes could vary dramatically. Thus every step of the research process should be seen as non-trivial because researcher variability occurs where we expect the highest levels of inter-researcher reliability: the generation of numerical results from an existing dataset with pre-specified independent and dependent variables.

Normally, we understand procedural bias as deriving from competencies or psychological predispositions in researchers that are activated when confronted with a multiplicity of data-analytic options(*10*, *11*) when traversing through the "garden of forking paths"(*8*). Differences in *competencies* are, for example, methodological and topical expertise and how these influence subjective judgments. *Confirmation bias* means that preexisting attitudes and beliefs lead researchers to make analytical choices that generate certain outcomes and justify these choices post-hoc. We suggest that idiosyncratic factors provide a third explanation independent of competencies or psychological biases. These are the more covert, potentially non-deliberate or un-measured(un-measurable) actions ineluctable in data analysis.

Identifying variability resulting from perverse incentives, model dependence and analytical robustness is nothing new(*9*, *10*, *12*, *13*), but scientists only recently began systematic assessment of the phenomenon by observing researchers in ecologically realistic yet controlled settings using 'many analysts' approaches. For instance, when 29 researchers tested if soccer referees were biased toward darker skin players using the same data, 29 unique model specifications were reported with empirical results ranging from modestly negative to strongly positive(*14*). Most of these studies were small in scale(*15*–*17*), aimed at investigating the reliability of narrow, field-specific methods(*15*, *18*, *19*) or were conducted in a research environment deliberately obtuse from involved researchers' substantive knowledge and foci(*14*). Here, we report findings from a large-scale crowdsourced research effort examining whether and to what degree variations in researchers' competencies, potential psychological biases and concomitant analytic steps affect the reliability of reported scientific findings using a common hypothesis and well-known international survey data.

Design

We (the principal investigators, "PIs": Breznau, Rinke, Wuttke) designed a study to make these devilishly-difficult-to-observe aspects of the analytical process observable: We coordinated a large group of data analysts ("participants") to simultaneously, yet independently, test a common and contested hypothesis in social science: whether immigration reduces support for social policies among the public(*20*).

Participants were given six survey questions to analyze opinions about different social policies from the *International Social Survey Programme*, a long-running, high-quality, multi-country survey widely used in the social sciences (see Communications in Supplementary Materials for sampling details). To remove potentially biasing incentives, all authors were ensured co-authorship regardless of their results, given preparatory tasks to familiarize them with the topic and asked to develop and pre-submit a research design to the PIs before running their tests. A total of 162 researchers in 73 teams submitted 1,261 models and 88 substantive conclusions (with some teams drawing two different conclusions). Their code was checked and then anonymized for public sharing by the PIs (Figs. S1,S2 and Tables S1,S2). Many teams submitted 12 models testing how two different immigration measures predict opinions about the six social policies (ranged from 1 to 124 models per team; mean=17.3).

Results

Fig. 1 visualizes our central insight: extensive variation in reported results.

Fig. 1 Broad variation in findings from 73 teams testing the same hypothesis with the same data

The distribution of estimated average marginal effects (AME) across all converged models (N = 1,253) includes results that are negative (green, and in the direction predicted by the given hypothesis the teams were testing), not different from zero (grey) or positive (orange), using a 95% confidence interval. AME are XY-standardized. Y-axis contains two breaks at +/-0.05. Numbers inside circles represent the percentage of the distribution of each outcome inversely weighted by the number of models per team.

We see the same pattern when we use the teams' conclusions rather than their statistical results: Different participants drew vastly different conclusions from the same set of data (see Figs. S5,S9,S10).

We find that competencies and potential confirmation biases do not explain the broad variation in outcomes.

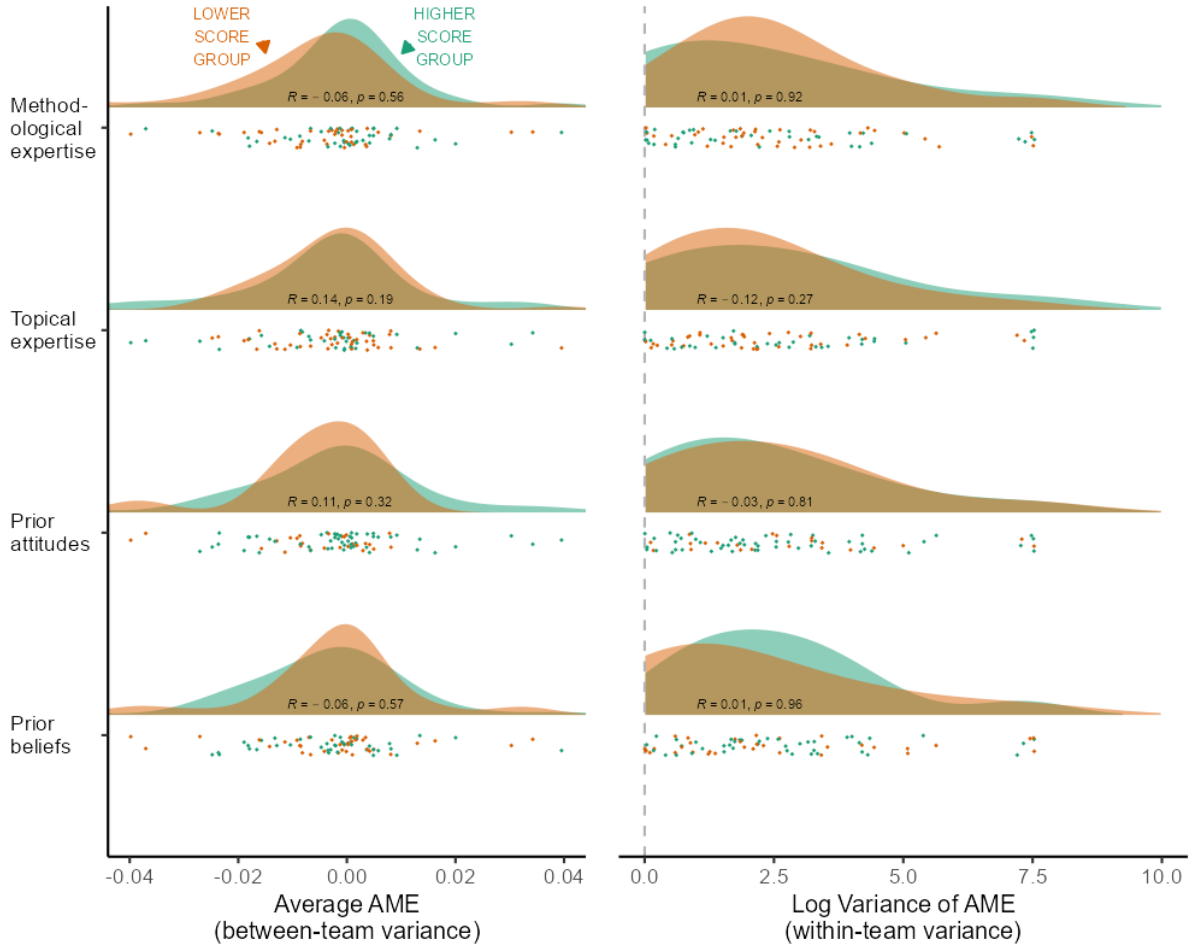

Fig. 2 Researcher characteristics do not explain outcome variance between teams or within teams

The distribution of team average of AMEs (left panel) and within-team variance in AMEs (right panel) across researchers grouped according to mean-splits ("LOWER" and "HIGHER") on methodological and topic expertise (potential competencies bias), and on prior attitudes toward immigration and beliefs about whether the hypothesis is true (potential confirmation bias). Log variance shifted so that minimum log value equals zero. Teams submitting only one model assigned a variance of zero. Pearson correlations along with a p-value ("R") are calculated using continuous scores of each researcher characteristic variable.

Researcher characteristics show no significant association with statistical results or even substantive conclusions. We confirmed this by hand-coding each research report submitted by the participants to identify their modeling choices. Examining all 1,261 models revealed 166 model specifications used by at least one team. "Specification" here means a single component of a model, for example measurement strategy, estimator, hierarchical structure, choice of independent variables and potential subsetting of the data. Of these 166, we found 107 were present in more than two teams' models, i.e., common

specifications. Most strikingly, the varying presence of these 107 specifications in a dissimilarity matrix revealed that no two models were an identical combination of them (Table S11).

Next we analyzed how much these specifications ("Research Design") of the participants could explain outcomes and substantive conclusions. Fig. 3 demonstrates that they mostly cannot.

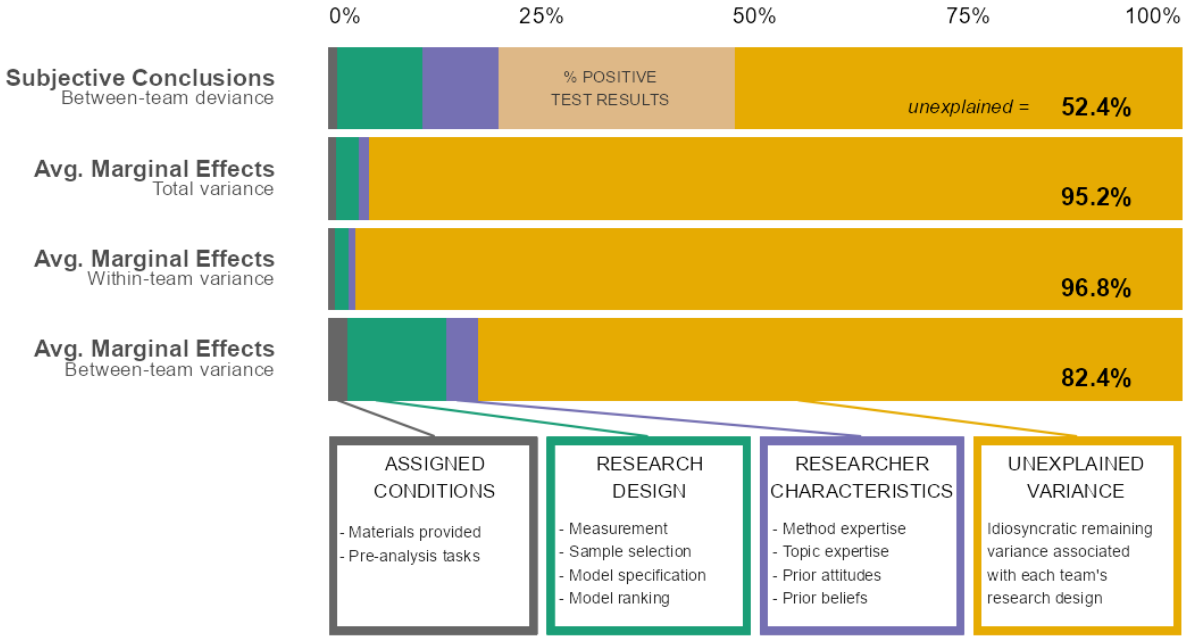

Fig. 3 Variance in statistical results and substantive conclusions between and within teams is mostly unexplained by conditions, research design and researcher characteristics
Decomposition of explained variance from multilevel general linear regression models using AMEs as the outcome (top three bars), and explained deviance from multinomial logistic regressions using the substantive conclusion(s) submitted by each team as a whole as to whether their results supported or rejected the hypothesis, or whether they thought the hypothesis was not testable given these data (bottom bar). We used informed step-wise addition of predictors to identify which specifications could explain the most variance/deviance while sacrificing the least degrees of freedom and maintaining the highest level of model fit based on log-likelihood and various information criteria; we also tested every possible combination as a robustness check. Equations, full regression results and definitions of all variables in "Main Regression Models" in Supplementary Materials.

The yellow portions of the explained deviance/variance bars dominate Fig. 3. Research design (2.6%, green segment) and researcher characteristics (1.2%, blue segment) explain almost none of the total variance (top bar) in the numerical results; a similar story whether we look at variance within or between teams. In other words, the massive variation in reported results originated from unique or idiosyncratic steps in the data analysis process leaving 95.2% of the total variance in results unexplained. Substantive conclusions are somewhat better explained, but only little. Research design (10.0%) and researcher characteristics (8.8%) explain only a portion of the deviance in substantive conclusions (bottom bar)

leaving 80.1% unexplained. Even the percentage of test results per team that statistically support their conclusions explain only about a third of the deviance (salmon-colored segment, bottom bar) leaving still 52.4% unexplained.

We confirmed the robustness of our results by automatically assessing every possible combination of model specifications. We also ran separate regressions by dependent variable and a variance function regression to check if specifications impacted the variability of results within teams and correct for potential heteroscedasticity (Tables S4,S9-S11). All results support the conclusion that a large part of the variance in research outcomes is from idiosyncratic researcher variability - unique analytical pathways vis-a-vis the rest of the participants.

To assess whether this unexplained variance can reasonably be considered 'surprising', we conducted a multiverse simulation suggesting that in a single workflow, model specifications can safely explain 16% of effect variance using the same data. This is way more than the 2.6% from research designs or 3.8% when including researcher characteristics, and it could be produced in a controlled setting using far fewer specifications (Table S8). Thus, further evidence that variations across workflows are unique.

Summary

The reliability of the results on which researchers base their conclusions is a central aspect of science to generate credible propositions and trustworthy solutions to problems. Much recent research focuses on systematic biases impacting reliability (publication bias, p-hacking, HARKing, etc.), yet we have little understanding of the more idiosyncratic sources of variation namely how much noise these sources introduce into the outcomes. Using a tightly controlled research design in a large-scale crowdsourced research effort involving 73 teams, we demonstrated that an ostensibly standardized and well-defined step in the research process – data analysis – can lead to substantial variation in statistical estimates and substantive conclusions.

We call this variation idiosyncratic because we removed incentives to arrive at any particular results and because we identified and adjusted for common model specification choices across teams. We adjusted for methods and topical expertise (potential competency bias) and for prior attitudes and beliefs (potential confirmation bias). We accounted for 107 model specifications. We were left with a surprising amount of unexplained variance: 95.2% of the total variance in AMEs and 80.1% of the deviance in substantive conclusions (52.4% if we include the percentage of positive test results in each team, see Fig. 3). In other words, even highly skilled scientists motivated to come to accurate results varied tremendously in what they found based on the same data and hypothesis.

These findings warrant discussion among those interested in meta-science and among the wider scientific community. If scientists are responsible for assessing and communicating uncertainty, they should address idiosyncratic variation. Their task is to measure a signal while attenuating noise as much as possible. Attenuation requires understanding the noise itself. Usually, this 'noise' is considered a product of instruments or alternative research designs. We conclude that if any given study had been conducted by a different (set of) researcher(s), perhaps even the same researchers at a different time, its results would likely have varied for reasons that cannot be reduced to easily identifiable analytical choices or biases.

The decisions made by researchers are so minute that they do not appear as decisions but rather as non-deliberate actions within the researcher's standard research setting. Yet, as we have shown, they are far from trivial. This means that every step should be held up to stronger theoretical scrutiny. Researchers should explicate the (implicit) theoretical assumptions associated with each step, for example dichotomizing a variable or choosing a specific estimator, or even choice of software if this were to introduce some non-trivial disturbance in the results.

Implications and limitations

We do not know the transmutability of our ecological setting to different topics, disciplines or even datasets. For instance, we cannot say how much researcher variability would impact the analysis of experimental data where the data-generating model is much clearer to the analyst. Moreover, in the social sciences there are no Newtonian laws or definite quantum statistical likelihoods to work with, suggesting our case might be less conservative than a similar study in the natural sciences. We also consider that the particular hypothesis we selected for this study has no consensus in the social science literature(*20*–*22*). It remains to be seen whether these data are more or less prone than other data to idiosyncratic variance in scientific results.

We see two implications for the current 'reproducibility crisis' narrative circulating across science. On the one hand, 15% of the teams had a majority of their statistical effects supporting the hypothesis (a negative estimate with a 95% CI, see Fig. S9) and 6.8% had a majority contradicting the hypothesis (positive/95% CI). This means that if publication is a rare event among a sample of researchers, random variation could easily lead to a universe of findings that are unambiguously in one direction or the other. If publication bias is a factor in this process(*23*), then the chance of this biased 'universe' of findings may increase simply as a product of noise in the research process. On the other hand, our findings suggest this ostensible crisis could be explained away as idiosyncratic variation inherent to the process of research, at least in theory. Thus, instead of a crisis, this would constitute a 'new opportunity and challenge' for scientific advancement(*24*). One such opportunity is to have even greater appreciation

for findings that have a high degree of consensus, for example on anthropogenic climate change or confirmation of Einstein's predictions from his Special Theory of Relativity.

Conclusion

Our study is the first to observe the mechanics of the data analysis process among social scientists at this level of detail. We find that a hidden source of noise may have major consequences for the reliability of scientific findings. If nothing else, scientists and consumers of their research should be humbler and cautious about the reliability of their findings. As we observed it here, idiosyncratic researcher variability is a phenomenon that should lead researchers to be more aware and embracing of the uncertainties permeating their work.

As our conclusions derive from myriad seemingly minor analytical decisions, just like the researchers we observed, we encourage readers to scrutinize these decisions as potentially non-trivial. Thus we provide replication files and an interactive tool ('Shiny app') to further explore the robustness checks presented in the Supplementary Materials document.

References

1. N. Oreskes, *Why Trust Science?* (Princeton University Press, Princeton, NJ, 2019).

2. M. Solomon, *Social Empiricism* (MIT press, Cambridge, MA, 2007).

3. C. F. Camerer, A. Dreber, F. Holzmeister, T.-H. Ho, J. Huber, M. Johannesson, M. Kirchler, G. Nave, B. A. Nosek, T. Pfeiffer, A. Altmejd, N. Buttrick, T. Chan, Y. Chen, E. Forsell, A. Gampa, E. Heikensten, L. Hummer, T. Imai, S. Isaksson, D. Manfredi, J. Rose, E.-J. Wagenmakers, H. Wu, Evaluating the replicability of social science experiments in Nature and Science between 2010 and 2015. *Nat. Hum. Behav.* **2**, 637–644 (2018).

4. Open Science Collaboration, Estimating the reproducibility of psychological science. *Science*. **349** (2015), doi:10.1126/science.aac4716.

5. S. Ritchie, *Science Fictions: How Fraud, Bias, Negligence, and Hype Undermine the Search for Truth* (Metropolitan Books, New York, NY, 2020).

6. A. B. Sørensen, The Structural basis of Social Inequality. *Am. J. Sociol.* **101**, 1333–1365 (1996).

7. B. S. Frey, Publishing as Prostitution? – Choosing Between One's Own Ideas and Academic Success. *Public Choice*. **116**, 205–223 (2003).

8. A. Gelman, E. Loken, The Statistical Crisis in Science. *Am. Sci.* **102**, 460 (2014).

9. J. P. Simmons, L. D. Nelson, U. Simonsohn, False-Positive Psychology: Undisclosed Flexibility in Data Collection and Analysis Allows Presenting Anything as Significant. *Psychol. Sci.* **22**, 1359–1366 (2011).

10. A. Orben, A. K. Przybylski, The association between adolescent well-being and digital technology use. *Nat. Hum. Behav.* (2019), doi:10.1038/s41562-018-0506-1.

11. M. Del Giudice, S. Gangestad, A Traveler's Guide to the Multiverse: Promises, Pitfalls, and a Framework for the Evaluation of Analytic Decisions. *Adv. Methods Pract. Psychol. Sci.* (2020).

12. U. Schimmack, A meta-psychological perspective on the decade of replication failures in social psychology. *Can. Psychol. Can.* **61**, 364–376 (2020).

13. J. Freese, D. Peterson, The Emergence of Statistical Objectivity: Changing Ideas of Epistemic Vice and Virtue in Science. *Sociol. Theory*. **36**, 289–313 (2018).

14. R. Silberzahn, E. L. Uhlmann, D. P. Martin, P. Anselmi, F. Aust, E. Awtrey, Š. Bahník, F. Bai, C. Bannard, E. Bonnier, R. Carlsson, F. Cheung, G. Christensen, R. Clay, M. A. Craig, A. Dalla Rosa, L. Dam, M. H. Evans, I. Flores Cervantes, N. Fong, M. Gamez-Djokic, A. Glenz, S. Gordon-McKeon, T. J. Heaton, K. Hederos, M. Heene, A. J. Hofelich Mohr, F. Högden, K. Hui, M. Johannesson, J. Kalodimos, E. Kaszubowski, D. M. Kennedy, R. Lei, T. A. Lindsay, S. Liverani, C. R. Madan, D. Molden, E. Molleman, R. D. Morey, L. B. Mulder, B. R. Nijstad, N. G. Pope, B. Pope, J. M. Prenoveau, F. Rink, E. Robusto, H. Roderique, A. Sandberg, E. Schlüter, F. D. Schönbrodt, M. F. Sherman, S. A. Sommer, K. Sotak, S. Spain, C. Spörlein, T. Stafford, L. Stefanutti, S. Tauber, J. Ullrich, M. Vianello, E.-J. Wagenmakers, M. Witkowiak, S. Yoon, B. A. Nosek, Many Analysts, One Data Set: Making Transparent How Variations in Analytic Choices Affect Results. *Adv. Methods Pract. Psychol. Sci.* **1**, 337–356 (2018).

15. G. Dutilh, J. Annis, S. D. Brown, P. Cassey, N. J. Evans, R. P. P. P. Grasman, G. E. Hawkins, A. Heathcote, W. R. Holmes, A.-M. Krypotos, C. N. Kupitz, F. P. Leite, V. Lerche, Y.-S. Lin, G. D.

Logan, T. J. Palmeri, J. J. Starns, J. S. Trueblood, L. van Maanen, D. van Ravenzwaaij, J. Vandekerckhove, I. Visser, A. Voss, C. N. White, T. V. Wiecki, J. Rieskamp, C. Donkin, The Quality of Response Time Data Inference: A Blinded, Collaborative Assessment of the Validity of Cognitive Models. *Psychon. Bull. Rev.* **26**, 1051–1069 (2019).

16. J. F. Landy, M. L. Jia, I. L. Ding, D. Viganola, W. Tierney, A. Dreber, M. Johannesson, T. Pfeiffer, C. R. Ebersole, Q. F. Gronau, Crowdsourcing hypothesis tests: Making transparent how design choices shape research results. *Psychol. Bull.* (2020).

17. J. J. Starns, A. M. Cataldo, C. M. Rotello, J. Annis, A. Aschenbrenner, A. Bröder, G. Cox, A. Criss, R. A. Curl, I. G. Dobbins, J. Dunn, T. Enam, N. J. Evans, S. Farrell, S. H. Fraundorf, S. D. Gronlund, A. Heathcote, D. W. Heck, J. L. Hicks, M. J. Huff, D. Kellen, K. N. Key, A. Kilic, K. C. Klauer, K. R. Kraemer, F. P. Leite, M. E. Lloyd, S. Malejka, A. Mason, R. M. McAdoo, I. M. McDonough, R. B. Michael, L. Mickes, E. Mizrak, D. P. Morgan, S. T. Mueller, A. Osth, A. Reynolds, T. M. Seale-Carlisle, H. Singmann, J. F. Sloane, A. M. Smith, G. Tillman, D. van Ravenzwaaij, C. T. Weidemann, G. L. Wells, C. N. White, J. Wilson, Assessing Theoretical Conclusions With Blinded Inference to Investigate a Potential Inference Crisis. *Adv. Methods Pract. Psychol. Sci.* **2**, 335–349 (2019).

18. J. A. Bastiaansen, Y. K. Kunkels, F. J. Blaauw, S. M. Boker, E. Ceulemans, M. Chen, S.-M. Chow, P. de Jonge, A. C. Emerencia, S. Epskamp, A. J. Fisher, E. L. Hamaker, P. Kuppens, W. Lutz, M. J. Meyer, R. Moulder, Z. Oravecz, H. Riese, J. Rubel, O. Ryan, M. N. Servaas, G. Sjobeck, E. Snippe, T. J. Trull, W. Tschacher, D. C. van der Veen, M. Wichers, P. K. Wood, W. C. Woods, A. G. C. Wright, C. J. Albers, L. F. Bringmann, Time to get personal? The impact of researchers choices on the selection of treatment targets using the experience sampling methodology. *J. Psychosom. Res.* **137**, 110211 (2020).

19. R. Botvinik-Nezer, F. Holzmeister, C. F. Camerer, A. Dreber, J. Huber, M. Johannesson, M. Kirchler, R. Iwanir, J. A. Mumford, R. A. Adcock, P. Avesani, B. M. Baczkowski, A. Bajracharya, L. Bakst, S. Ball, M. Barilari, N. Bault, D. Beaton, J. Beitner, R. G. Benoit, R. M. W. J. Berkers, J. P. Bhanji, B. B. Biswal, S. Bobadilla-Suarez, T. Bortolini, K. L. Bottenhorn, A. Bowring, S. Braem, H. R. Brooks, E. G. Brudner, C. B. Calderon, J. A. Camilleri, J. J. Castrellon, L. Cecchetti, E. C. Cieslik, Z. J. Cole, O. Collignon, R. W. Cox, W. A. Cunningham, S. Czoschke, K. Dadi, C. P. Davis, A. D. Luca, M. R. Delgado, L. Demetriou, J. B. Dennison, X. Di, E. W. Dickie, E. Dobryakova, C. L. Donnat, J. Dukart, N. W. Duncan, J. Durnez, A. Eed, S. B. Eickhoff, A. Erhart, L. Fontanesi, G. M. Fricke, S. Fu, A. Galván, R. Gau, S. Genon, T. Glatard, E. Glerean, J. J. Goeman, S. A. E. Golowin, C. González-García, K. J. Gorgolewski, C. L. Grady, M. A. Green, J. F. Guassi Moreira, O. Guest, S. Hakimi, J. P. Hamilton, R. Hancock, G. Handjaras, B. B. Harry, C. Hawco, P. Herholz, G. Herman, S. Heunis, F. Hoffstaedter, J. Hogeveen, S. Holmes, C.-P. Hu, S. A. Huettel, M. E. Hughes, V. Iacovella, A. D. Iordan, P. M. Isager, A. I. Isik, A. Jahn, M. R. Johnson, T. Johnstone, M. J. E. Joseph, A. C. Juliano, J. W. Kable, M. Kassinopoulos, C. Koba, X.-Z. Kong, T. R. Koscik, N. E. Kucukboyaci, B. A. Kuhl, S. Kupek, A. R. Laird, C. Lamm, R. Langner, N. Lauharatanahirun, H. Lee, S. Lee, A. Leemans, A. Leo, E. Lesage, F. Li, M. Y. C. Li, P. C. Lim, E. N. Lintz, S. W. Liphardt, A. B. Losecaat Vermeer, B. C. Love, M. L. Mack, N. Malpica, T. Marins, C. Maumet, K. McDonald, J. T. McGuire, H. Melero, A. S. Méndez Leal, B. Meyer, K. N. Meyer, G. Mihai, G. D. Mitsis, J. Moll, D. M. Nielson, G. Nilsonne, M. P. Notter, E. Olivetti, A. I. Onicas, P. Papale, K. R. Patil, J. E. Peelle, A. Pérez, D. Pischedda, J.-B. Poline, Y. Prystauka, S. Ray, P. A. Reuter-Lorenz, R. C. Reynolds, E. Ricciardi, J. R. Rieck, A. M. Rodriguez-Thompson, A. Romyn, T. Salo, G. R. Samanez-Larkin, E. Sanz-Morales, M. L. Schlichting, D. H. Schultz, Q. Shen, M. A. Sheridan, J. A. Silvers, K. Skagerlund, A. Smith, D. V. Smith, P. Sokol-Hessner, S. R. Steinkamp, S. M. Tashjian, B. Thirion, J. N. Thorp, G. Tinghög, L. Tisdall, S. H. Tompson, C. Toro-Serey, J. J. Torre Tresols, L. Tozzi, V. Truong, L. Turella, A. E. van ‘t Veer, T. Verguts, J. M. Vettel, S. Vijayarajah, K. Vo, M. B. Wall, W. D. Weeda, S. Weis, D. J. White, D. Wisniewski, A. Xifra-Porxas, E. A. Yearling, S. Yoon, R. Yuan, K. S. L. Yuen,

L. Zhang, X. Zhang, J. E. Zosky, T. E. Nichols, R. A. Poldrack, T. Schonberg, Variability in the analysis of a single neuroimaging dataset by many teams. *Nature*. **582**, 84–88 (2020).

20. D. Brady, R. Finnigan, Does Immigration Undermine Public Support for Social Policy? *Am. Sociol. Rev.* **79**, 17–42 (2014).

21. M. A. Eger, N. Breznau, Immigration and the Welfare State: A Cross-Regional Analysis of European Welfare Attitudes. *Int. J. Comp. Sociol.* **58**, 440–463 (2017).

22. A. Alesina, S. Stantcheva, E. Teso, Intergenerational Mobility and Preferences for Redistribution. *Am. Econ. Rev.* **108**, 521–554 (2018).

23. A. Franco, N. Malhotra, G. Simonovits, Publication bias in the social sciences: Unlocking the file drawer. *Science*. **345**, 1502–1505 (2014).

24. D. Fanelli, Opinion: Is science really facing a reproducibility crisis, and do we need it to? *Proc. Natl. Acad. Sci.* **115**, 2628–2631 (2018).